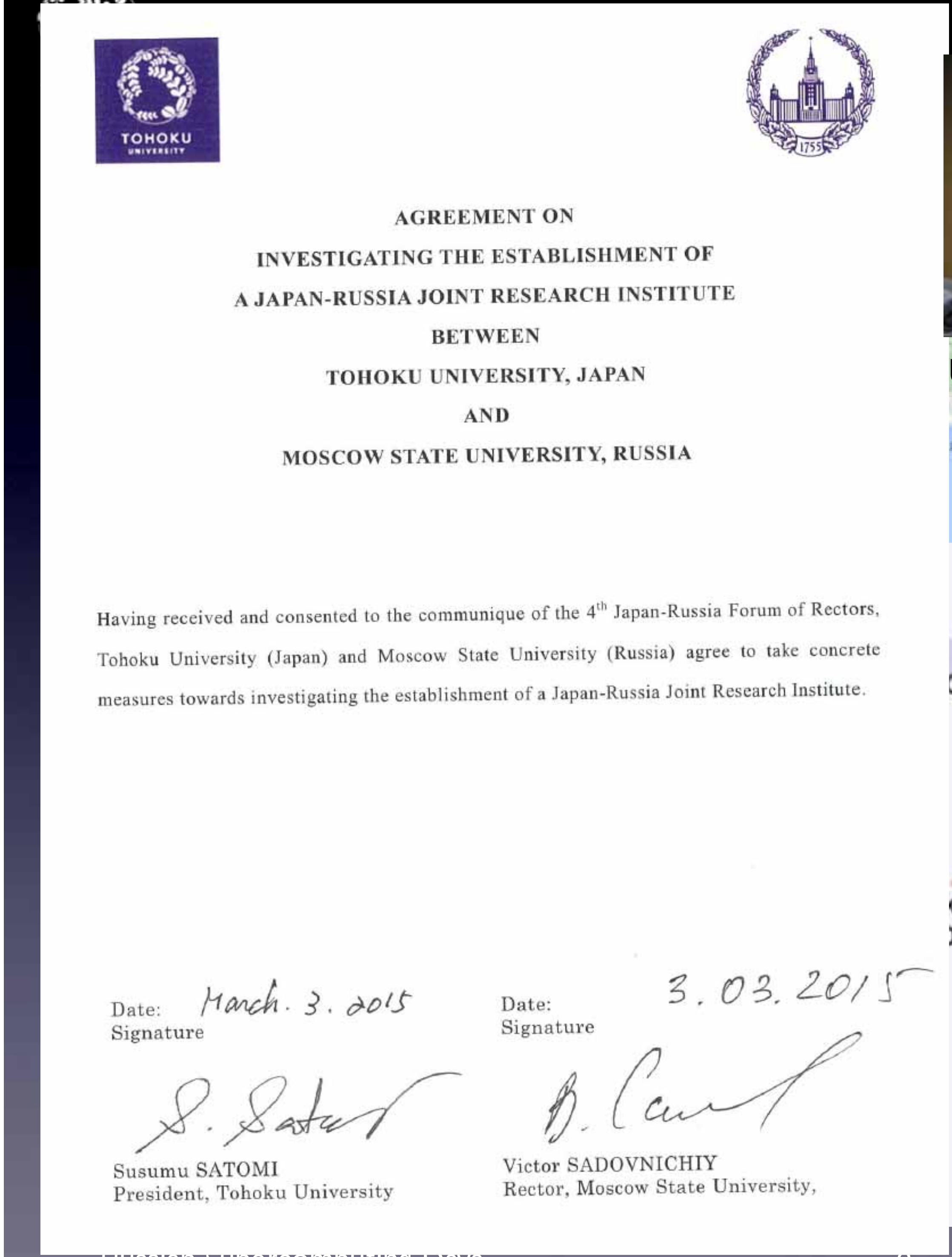
Highly Productive HPC on Modern Vector Supercomputers: Current and Future

Hiroaki Kobayashi
Director and Professor
Cyberscience Center
Tohoku University
koba@cc.tohoku.ac.jp

Russian Supercomputing Days
Moscow, Russia
September 28-29, 2015



TOHOKU UNIVERSITY

AGREEMENT ON
INVESTIGATING THE ESTABLISHMENT OF
A JAPAN-RUSSIA JOINT RESEARCH INSTITUTE
BETWEEN
TOHOKU UNIVERSITY, JAPAN
AND
MOSCOW STATE UNIVERSITY, RUSSIA

Having received and consented to the communique of the 4th Japan-Russia Forum of Rectors, Tohoku University (Japan) and Moscow State University (Russia) agree to take concrete measures towards investigating the establishment of a Japan-Russia Joint Research Institute.

Date: March. 3. 2015
Signature

Susumu SATOMI
President, Tohoku University

Date: 3.03.2015
Signature

Victor SADOVNICHIY
Rector, Moscow State University,

underg

Missions of Cyberscience Center
As a National Supercomputer Center

★ High-Performance Computing Center founded in 1969

- Offering leading-edge high-performance computing environments to academic users nationwide in Japan
  - 24/7 operations of large-scale vector-parallel and scalar-parallel systems
  - 1500 users registered in AY 2014
- User supports
  - Benchmarking, analyzing, and tuning users' programs
  - Holding seminars and lectures
- Supercomputing R&D, collaborating work with NEC
  - Designing next-generation high-performance computing systems and their applications for highly-productive supercomputing
  - 57-year history of collaboration between Tohoku University and NEC on High Performance Vector Computing
- Education
  - Teaching and supervising BS, MS and Ph.D. Students as a cooperative laboratory of graduate school of information sciences, Tohoku university

1969

1982

SX-1 in 1985

SX-2 in 1989

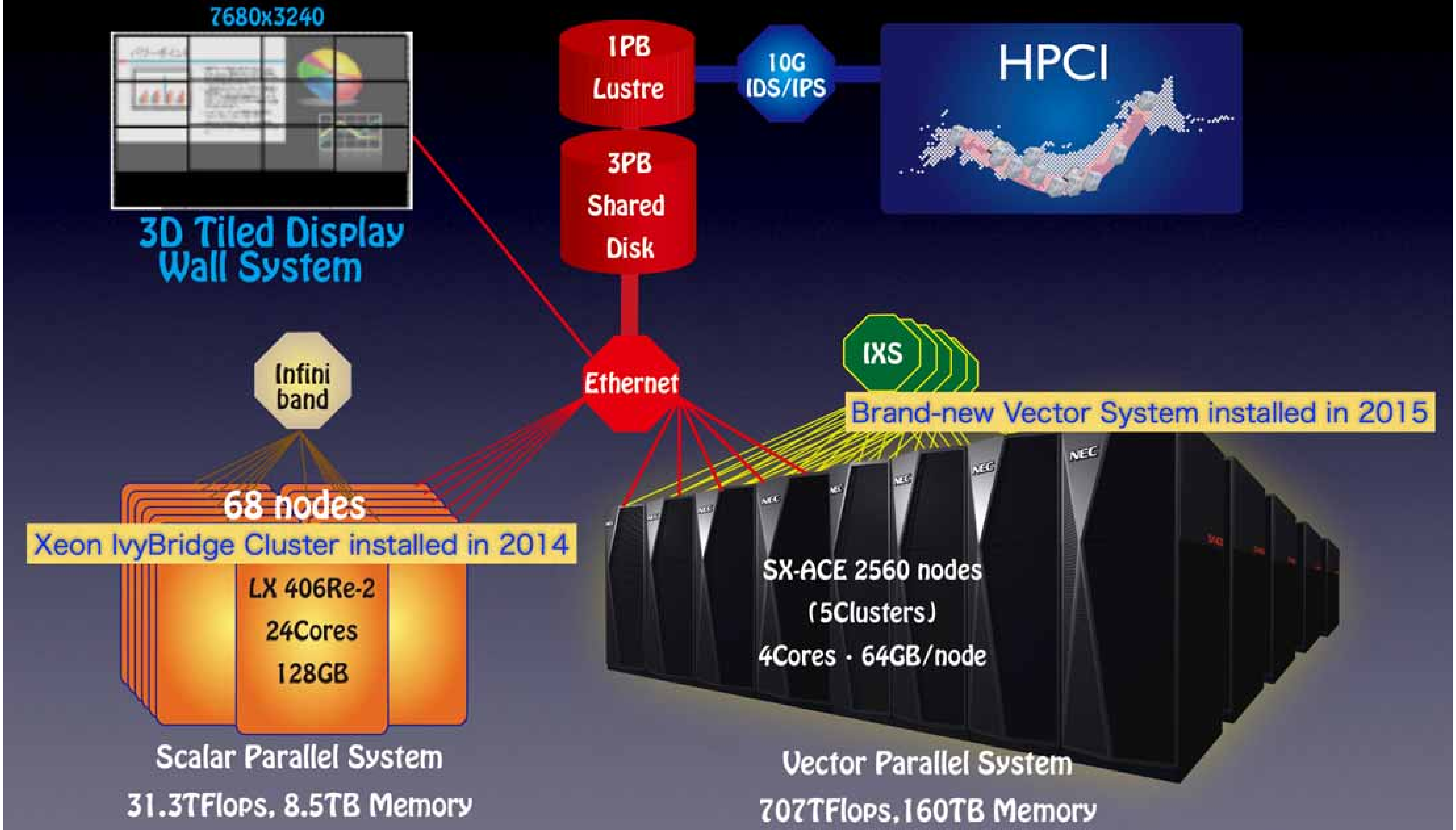
SX-3 in 1994

SX-4 in 1998

SX-7 in 2003

SX-9 in 2008

Tohoku Univ.'s New Supercomputer System (2015.2.20~)



7680x3240

3D Tiled Display Wall System

1PB Lustre

10G IDS/IPS

HPCI

3PB Shared Disk

Ethernet

Infini band

IXS

Brand-new Vector System installed in 2015

68 nodes

Xeon IvyBridge Cluster installed in 2014

LX 406Re-2

24Cores

128GB

SX-ACE 2560 nodes

(5Clusters)

4Cores · 64GB/node

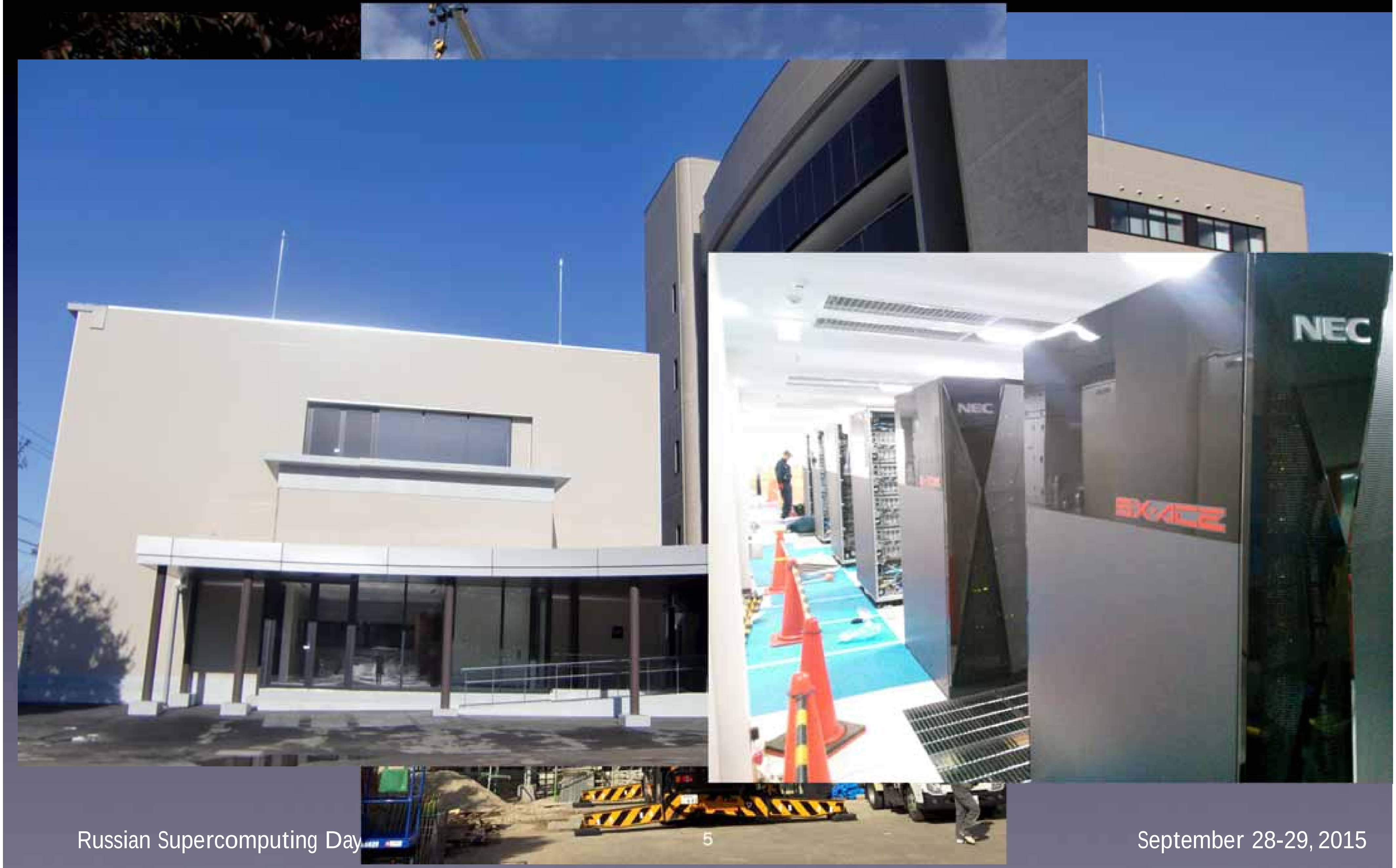
Scalar Parallel System

31.3TFlops, 8.5TB Memory

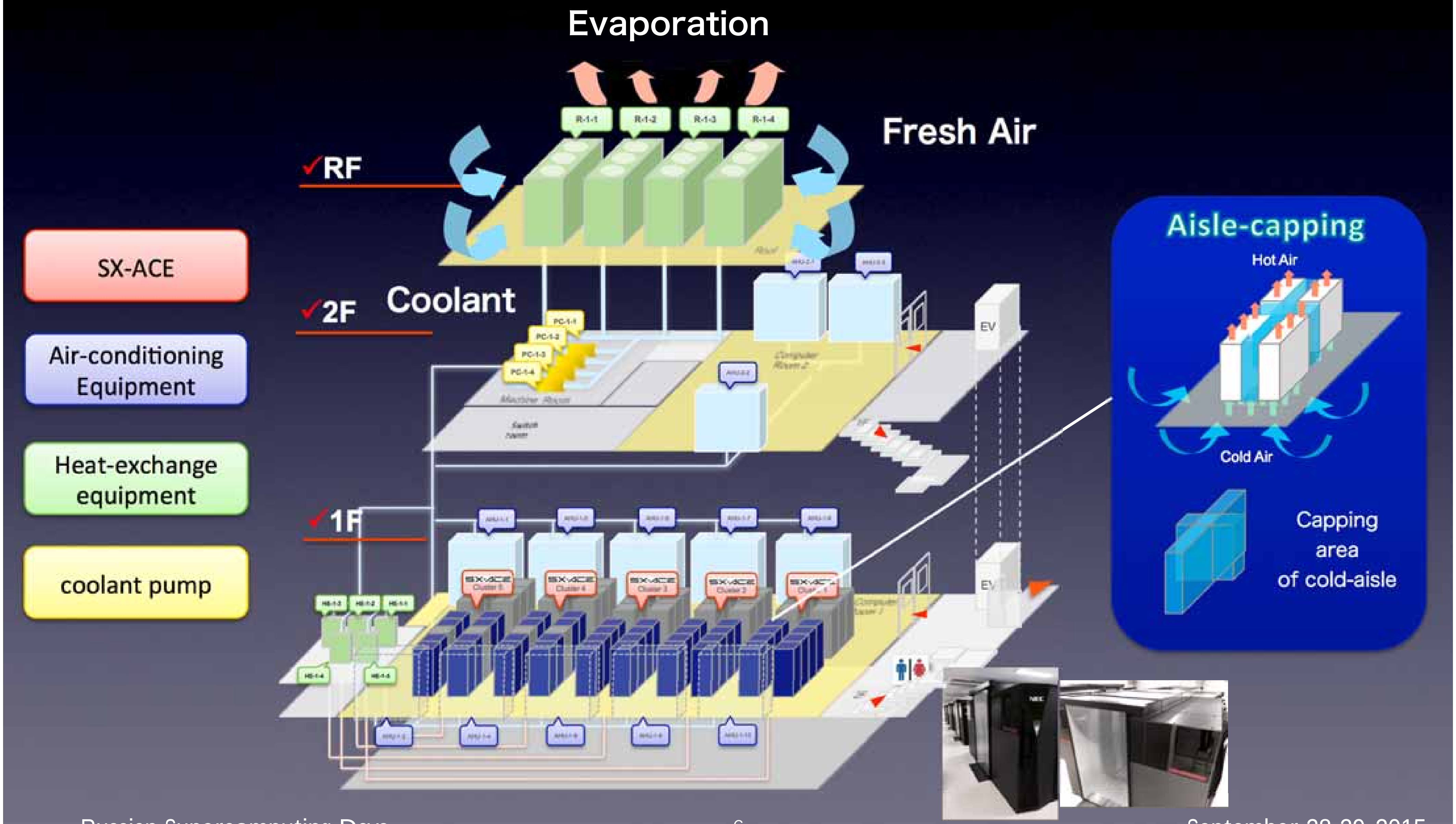
Vector Parallel System

707TFlops,160TB Memory

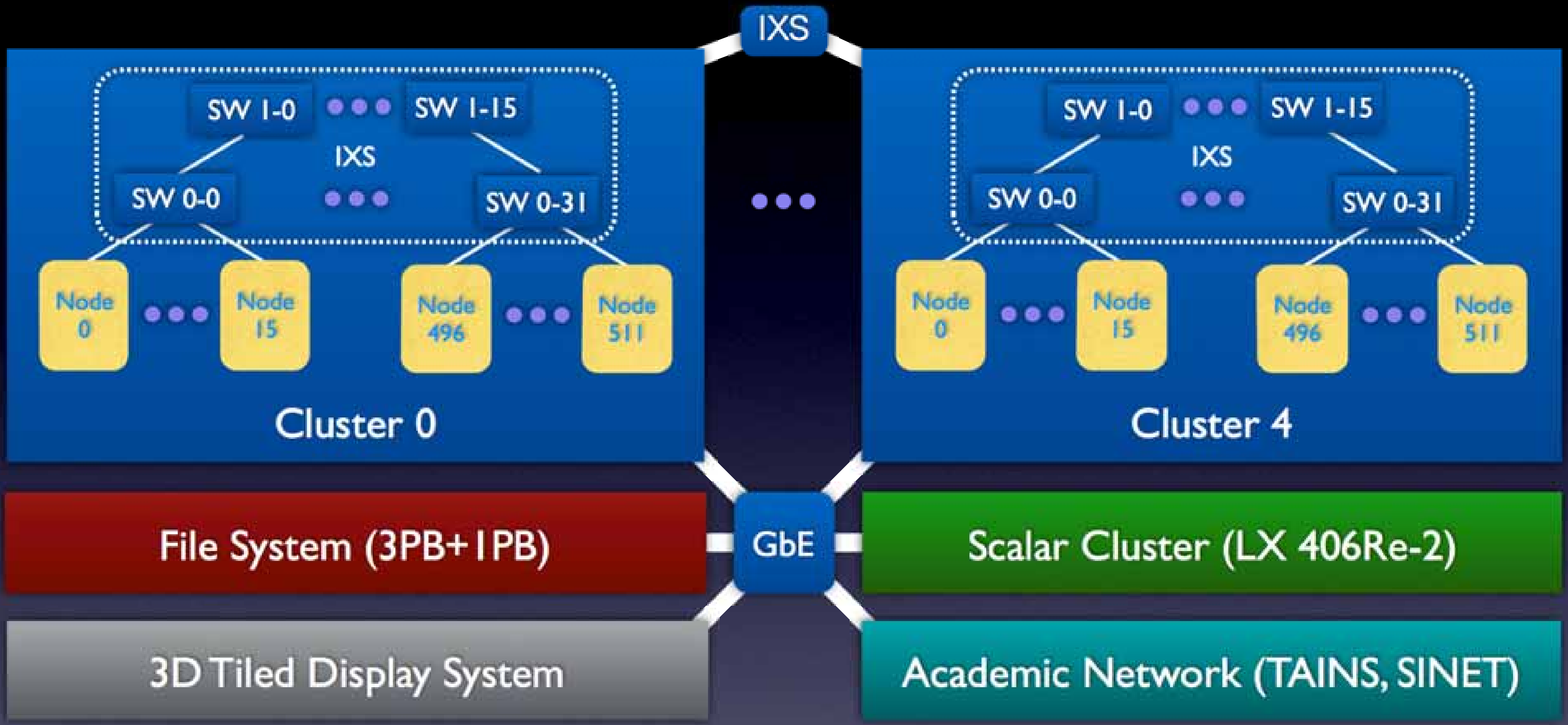
New HPC Building Construction and System Installation (2014.7~2015.2)



Cooling Facility of HPC Building



Organization of Tohoku Univ. SX-ACE System



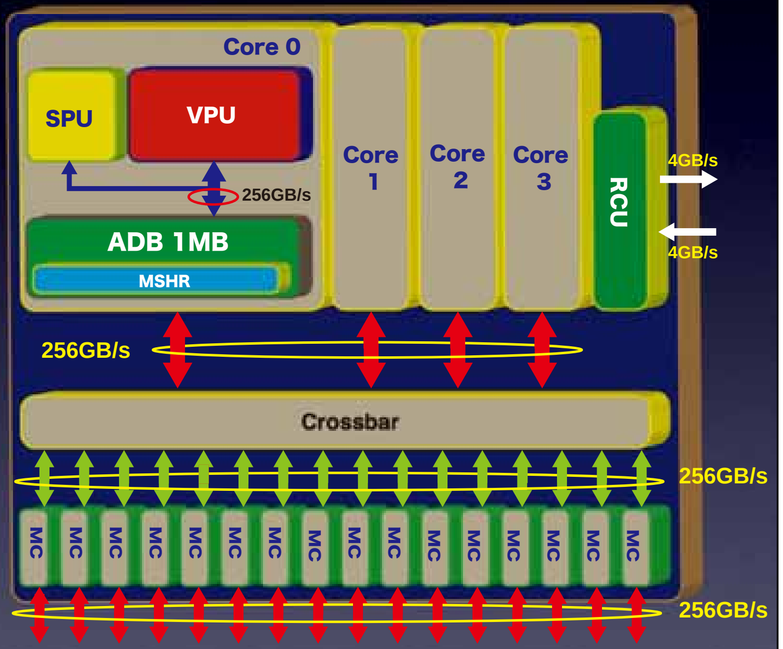
	Core	CPU(Socket)	Node	Cluster	Total System
Size	1	4 Cores	1CPU	512 Nodes	5 Clusters
Performance (VPU+SPU)	69GFlop/s (68GF+1GF)	276GFlop/s (272GF+4GF)		141Tflop/s (139TF+ 2TF)	707Tflop/s (697TF+10TF)
Mem. BW	256GB/s			131TB/s	655TB/s
Memory Cap.	64GB			32TB	160TB
IXS Node BW	-		4GB/s x2		-

Features of the SX-ACE Vector Processor

- 4 Core Configuration, each with High-Performance Vector-Processing Unit and Scalar Processing Unit
  - 272Gflop/s of VPU + 4Gflop/s of SPU per socket
    - 68Gflop/s + 1Gflop/s per core
  - 1MB private ADB per core (4MB per socket)
    - Software-controlled on-chip memory for vector load/store
    - 4x compared with SX-9
    - 4-way set-associative
    - MSHR with 512 entries (address+data)
    - 256GB/s to/from Vec. Reg.
      - 4B/F for Multiply-Add operations
  - 256 GB/s memory bandwidth, Shared with 4 cores
    - 1B/F in 4-core Multiply-Add operations

      ～ 4B/F in 1-core Multiply-Add operations
    - 128 memory banks per socket
- Other improvement and new mechanisms to enhance vector processing capability, especially for efficient handling of short vectors operations and indirect memory accesses
  - Out of Order execution for vector load/store operations
  - Advanced data forwarding in vector pipes chaining
  - Shorter memory latency than SX-9

SX-ACE Processor Architecture



Source: NEC

Features of Tohoku Univ. SX-ACE System

Significant Performance Improvement with Lower Power and Less Space

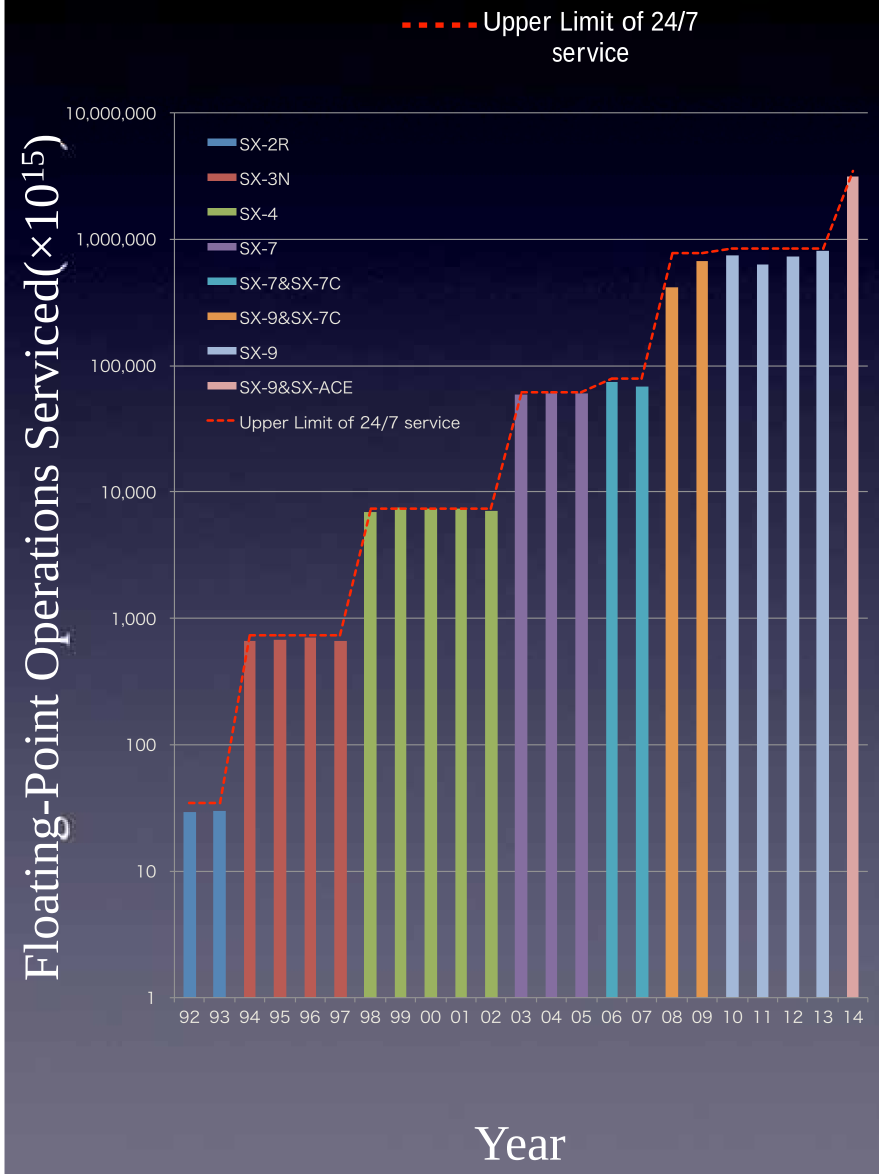
		SX-9 (2008)	SX-ACE (2014)	Improvement
	Number of Cores	1	4	4x
CPU Performance	Total Flop/s	118.4Gflop/s	276Gflop/s	2.3x
	Memory Bandwidth	256GB/sec	256GB/sec	1
	ADB Capacity	256KB	4MB	16x
Total Performance, Footprint, Power Consumption	Total Flop/s	34.1Tfop/s	706.6Tflop/s	20.7x
	Total Memory Bandwidth	73.7TB/s	655TB/s	8.9x
	Total Memory Capacity	18TB	160TB	8.9x
	Power Consumption	590kVA	1,080kVA	1.8x
	Footprint	**$293m^2$**	**$430m^2$**	1.5x

Powerful CPU/Node Performance and Higher B/F rate

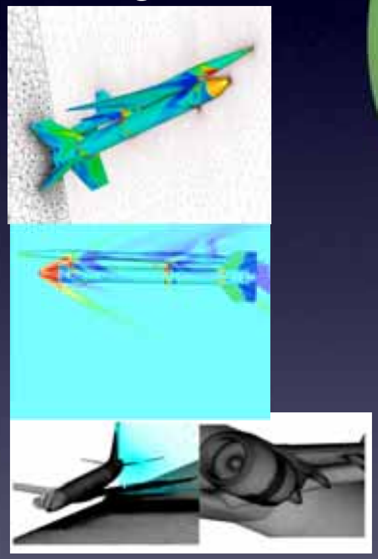
		SX-ACE(2014)	**K(2011)**	**Ratio**
CPU (Node) Performance	Clock Frequency	1GHz	2GHz	0.5x
	Flop/s per Core	64Gflop/s	16Gflop/s	4x
	Cores per CPU	4	8	0.5x
	Flop/s per CPU	256Gflop/s	128Gflop/s	2x
	Bandwidth	256GB/s	64GB/s	4x
	Bytes per Flop (B/F)	1	0.5	2x
	Memory Capacity	64GB	16GB	4x

A Balanced System for High Sustained Performance, resulting in High Productivity in the Wide Area of Applications in Academia and Industry

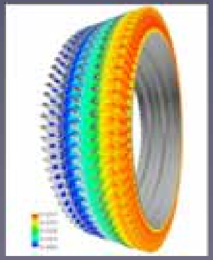
High Demands for Vector Systems in Memory-Intensive, Science and Engineering Applications

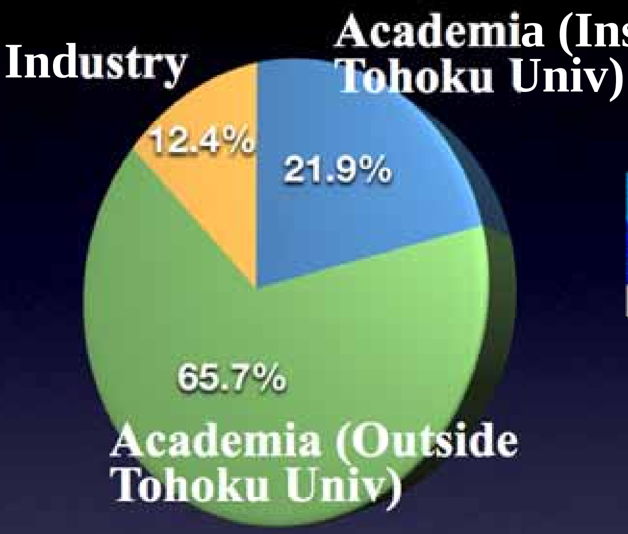


Advanced CFD model for digital flight

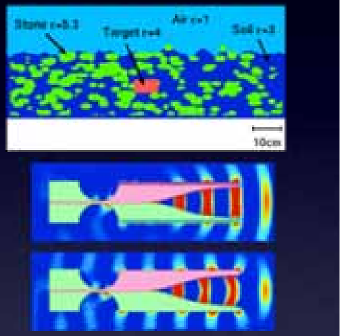


Turbine Design



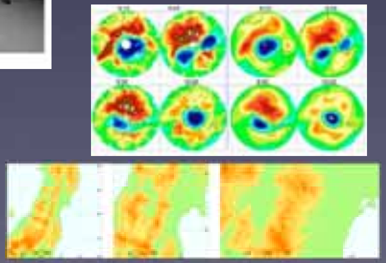


Electromagnetic Field Analysis

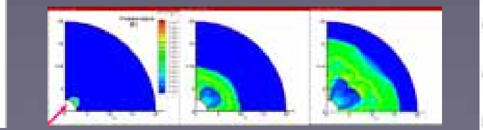


Catalyzer Design

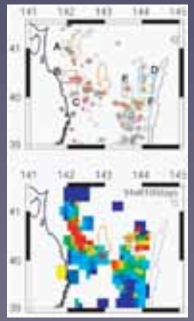
Atmosphere Analysis& Whether Forecasting



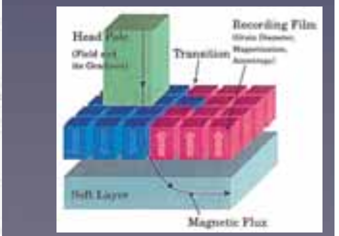
Combustion



Earthquake analysis



Magnetic recoding device Design

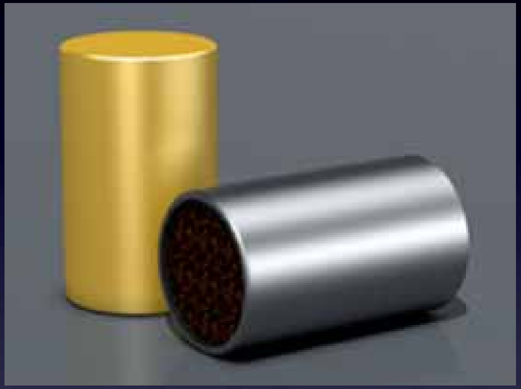


Expanding Industrial Use

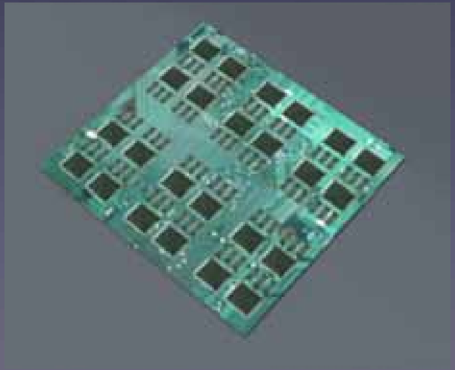
TV program “Close-up GENDAI” by NHK (2013.1.8)

Advanced Perpendicular Magnetic Recording Hard Drive

Highly Efficient Turbines for Power Plants

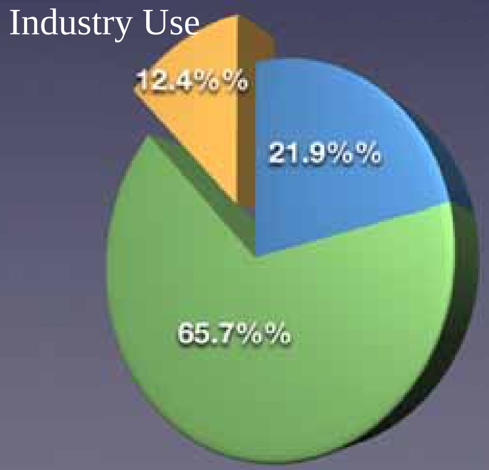


Exhaust Gas Catalyst



Base Material for PCBs

Regional Jet



Performance Evaluations of SX-ACE

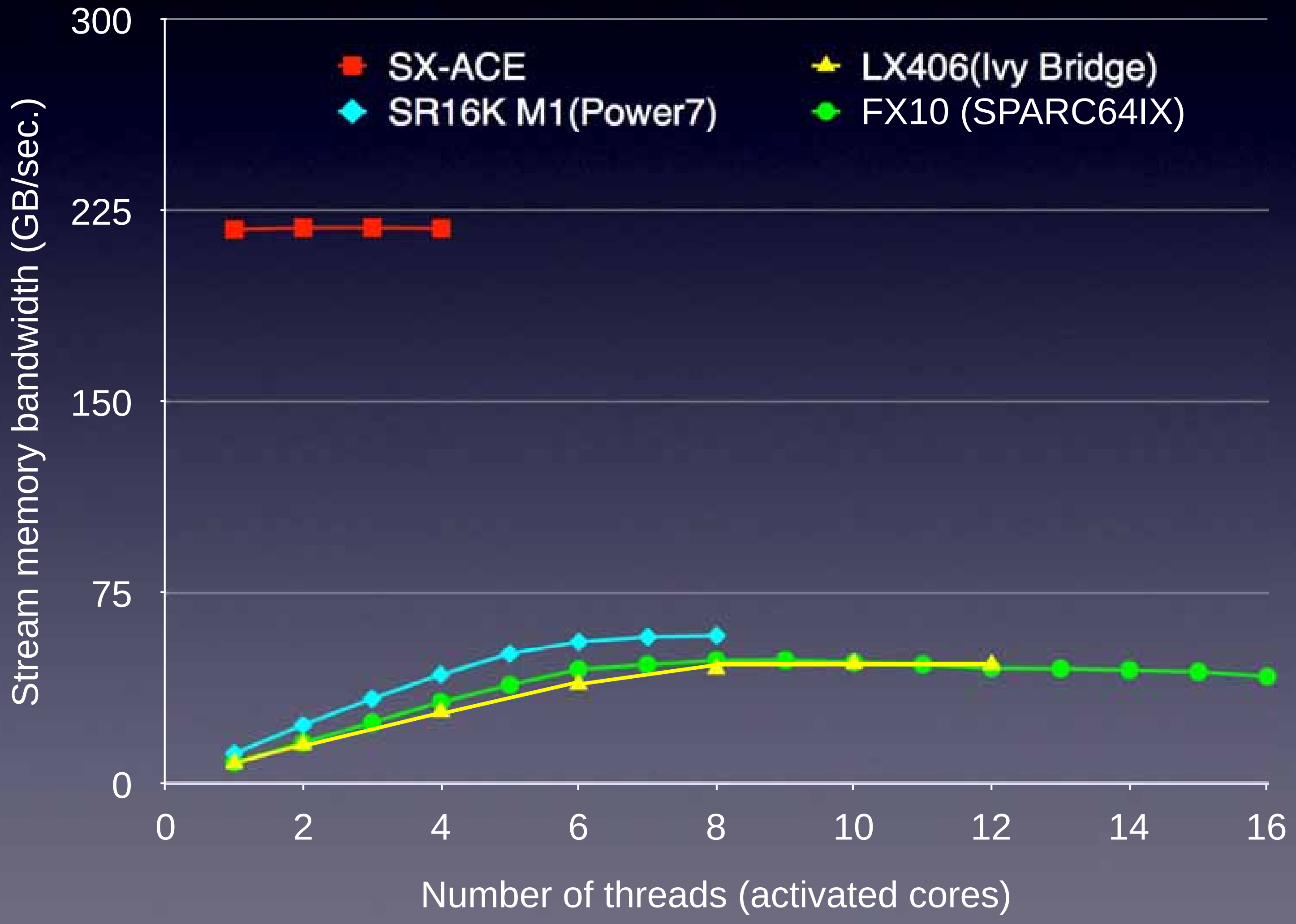
Specifications of Evaluated Systems

System	No. of Sockets/ Node	Perf./ Socket (Gflop/s)	No. of Cores	Perf. /core (Gflop/s)	Mem. BW GB/sec	On-chip mem	NW BW (GB/sec)	Sys. B/F
SX-ACE	1	256	4	64	256	1MB ADB /core	2 x 4 IXS	1.0
SX-9	16	102.4	1	102.4	256	256KB ADB/core	2 x 128 IXS	2.5
ES2	8	102.4	1	102.4	256	256KB ADB/core	2 x 64IXS	2.5
LX 406 (Ivy Bridge)	2	230.4	12	19.2	59.7	256KB L2/core 30MB Shared L3	5 IB	0.26
FX10 (SPARK64IX)	1	236.5	16	14.78	85	12MB shared L2	5 - 50 Tofu NW	0.36
K (SPARK64VIII)	1	128	8	16	64	6MB Shared L2	5 - 50 Tofu NW	0.5
SR16K M1 (Power7)	4	245.1	8	30.6	128	256KB L2/core 32MB shared L3	2 x 24 - 96 custom NW	0.52

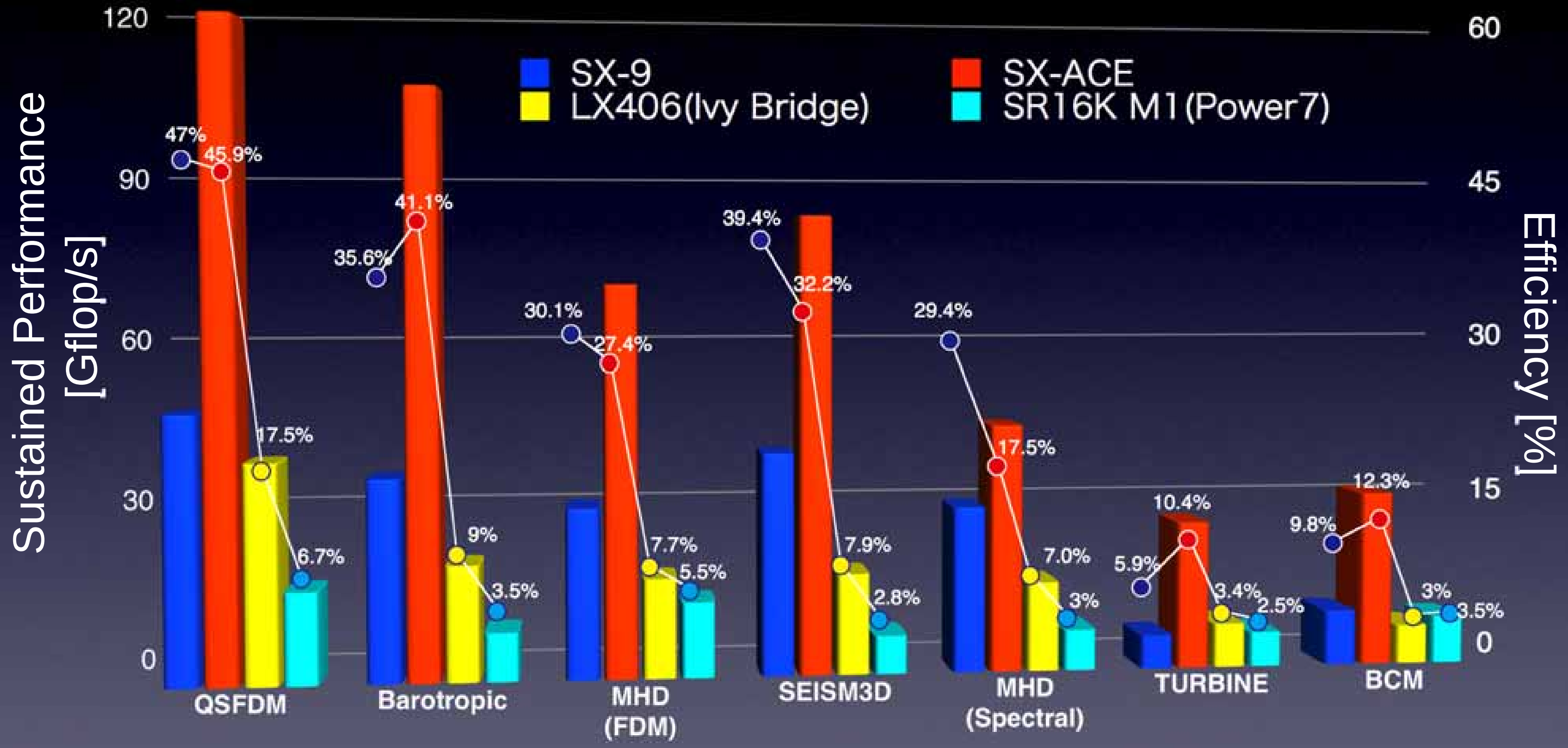
Remarks: Listed performances are obtained based on total Multiply-Add performances of individual systems

Sustained Memory Bandwidth

- STREAM (TRIAD)

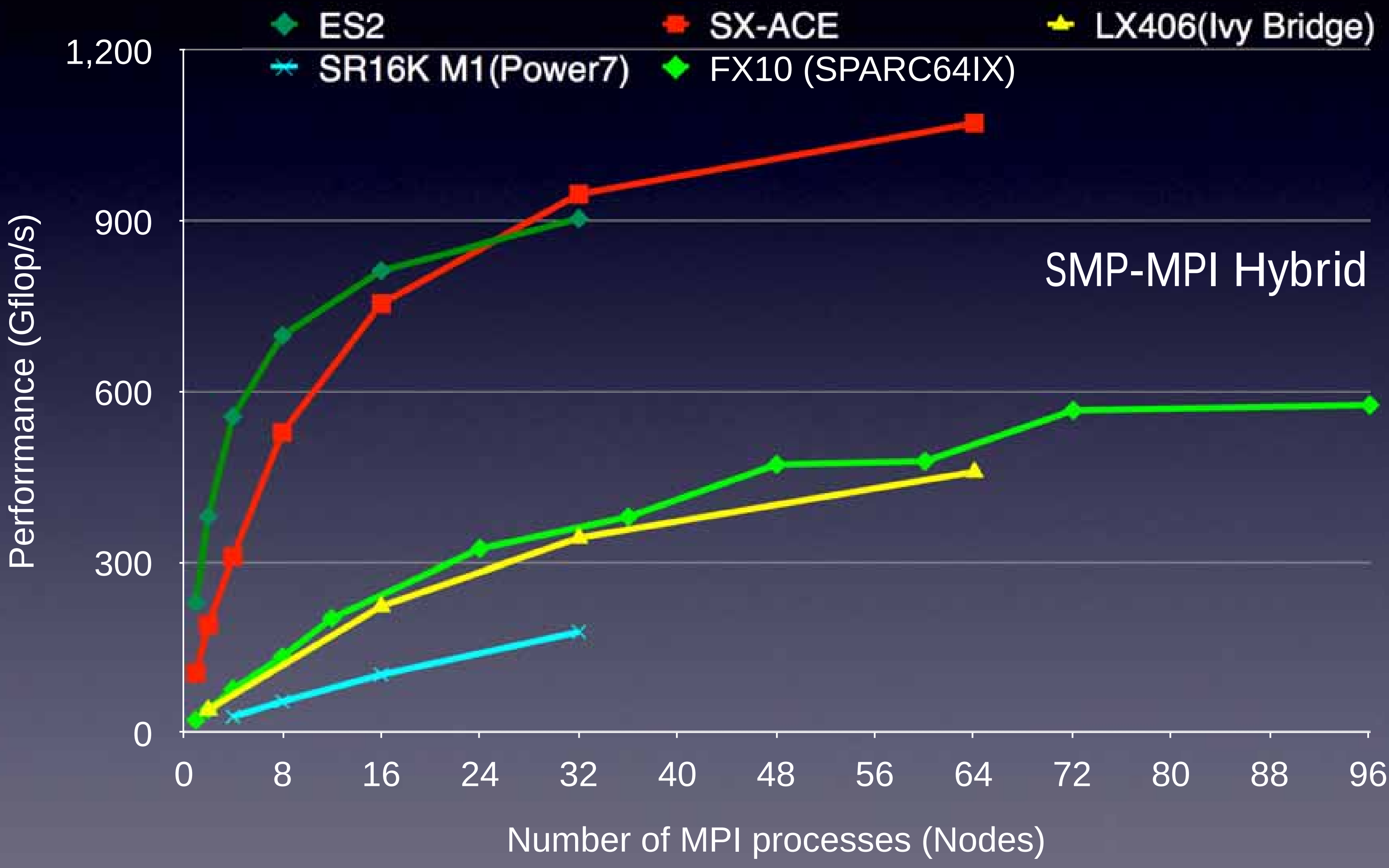


Sustained Single CPU Performance



	QSFDM	Barotropic	MHD (FDM)	Seism3D	MHD (Spectral)	TURBINE	BCM
Code B/F Memory Intensity	2.16	1.97	3.04	2.15	2.21	1.78	7.01

Sustained Performance of Barotropic Ocean Model on Multi-Node Systems



Performance Evaluation of SX-ACE by using HPCG

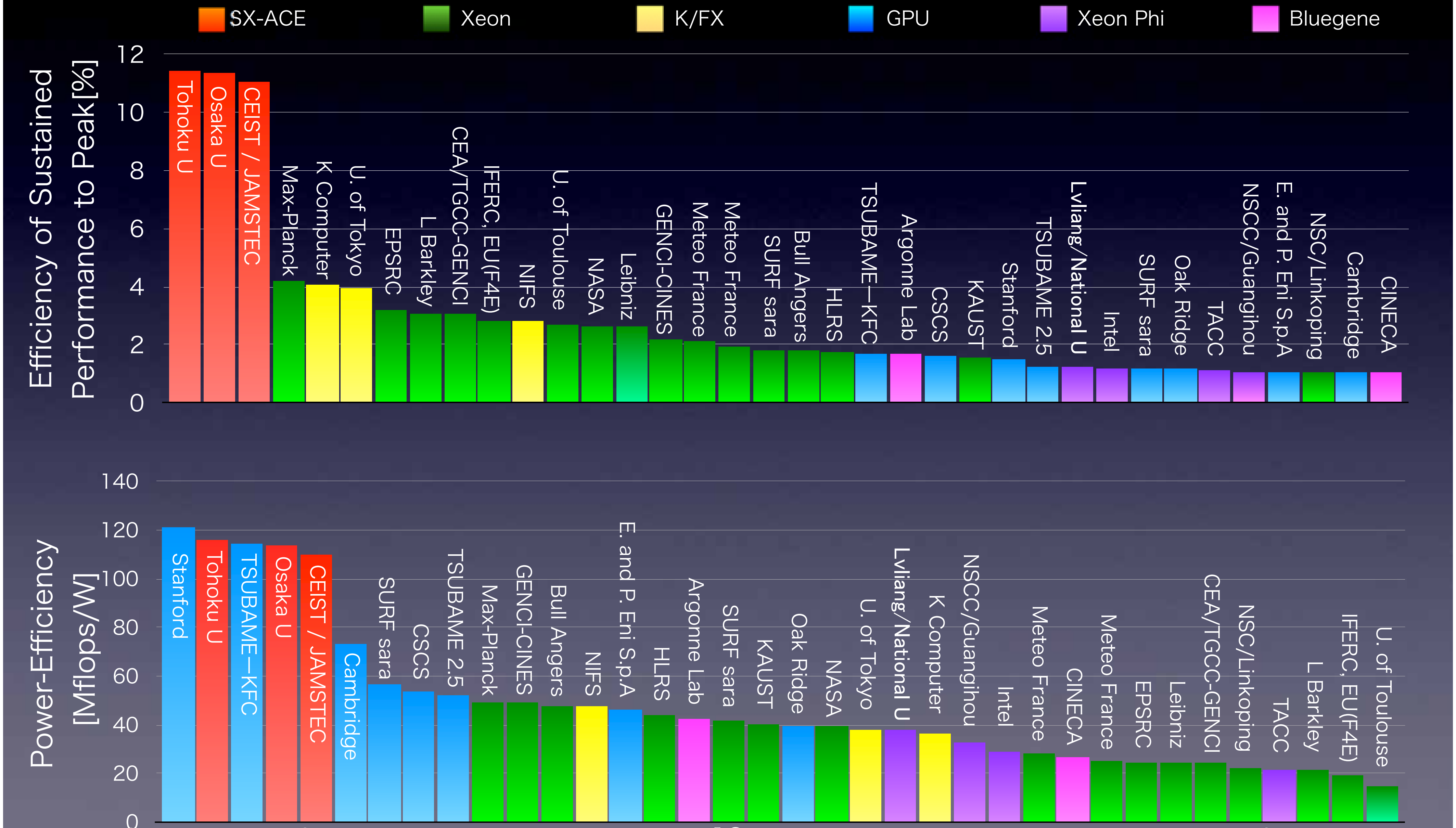
★ HPCG (High Performance Conjugate Gradients) is designed

- to exercise computational and data access patterns that more closely match a broad set of important applications, and
- to give incentive to computer system designers to invest in capabilities that will have impact on the collective performance of these applications.
  - ✓ HPL for top500 is increasingly unreliable as a true measure of system performance for a growing collection of important science and engineering applications.

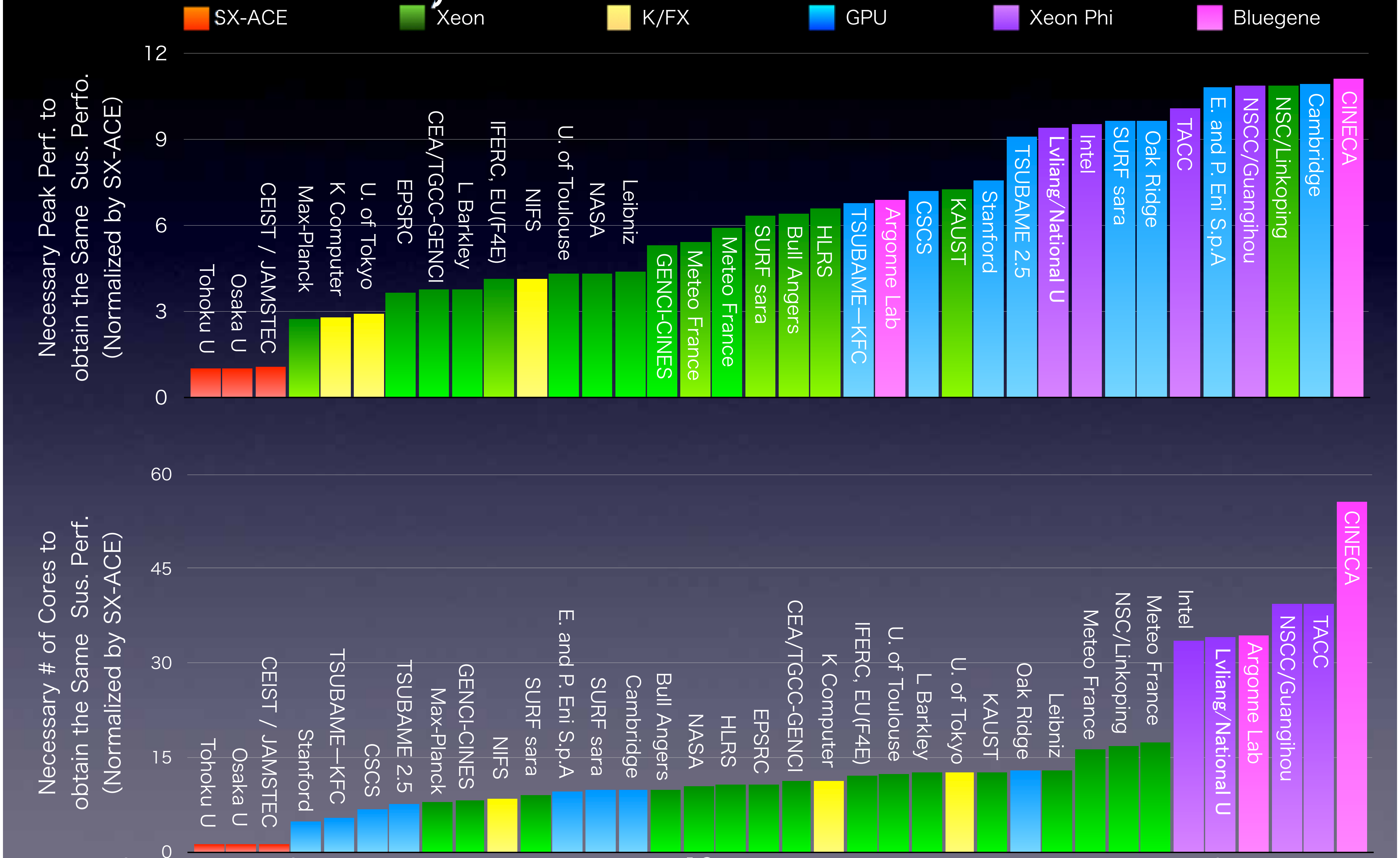
★ HPCG is a complete, stand-alone code that measures the performance of basic operations in a unified code:

- ✓ Sparse matrix-vector multiplication.
- ✓ Sparse triangular solve.
- ✓ Vector updates.
- ✓ Global dot products.
- ✓ Local symmetric Gauss-Seidel smoother.
- ✓ Driven by multigrid preconditioned conjugate gradient algorithm that exercises the key kernels on a nested set of coarse grids.
- ✓ Reference implementation is written in C++ with MPI and OpenMP support.

Efficiency Evaluation of HPCG Performance

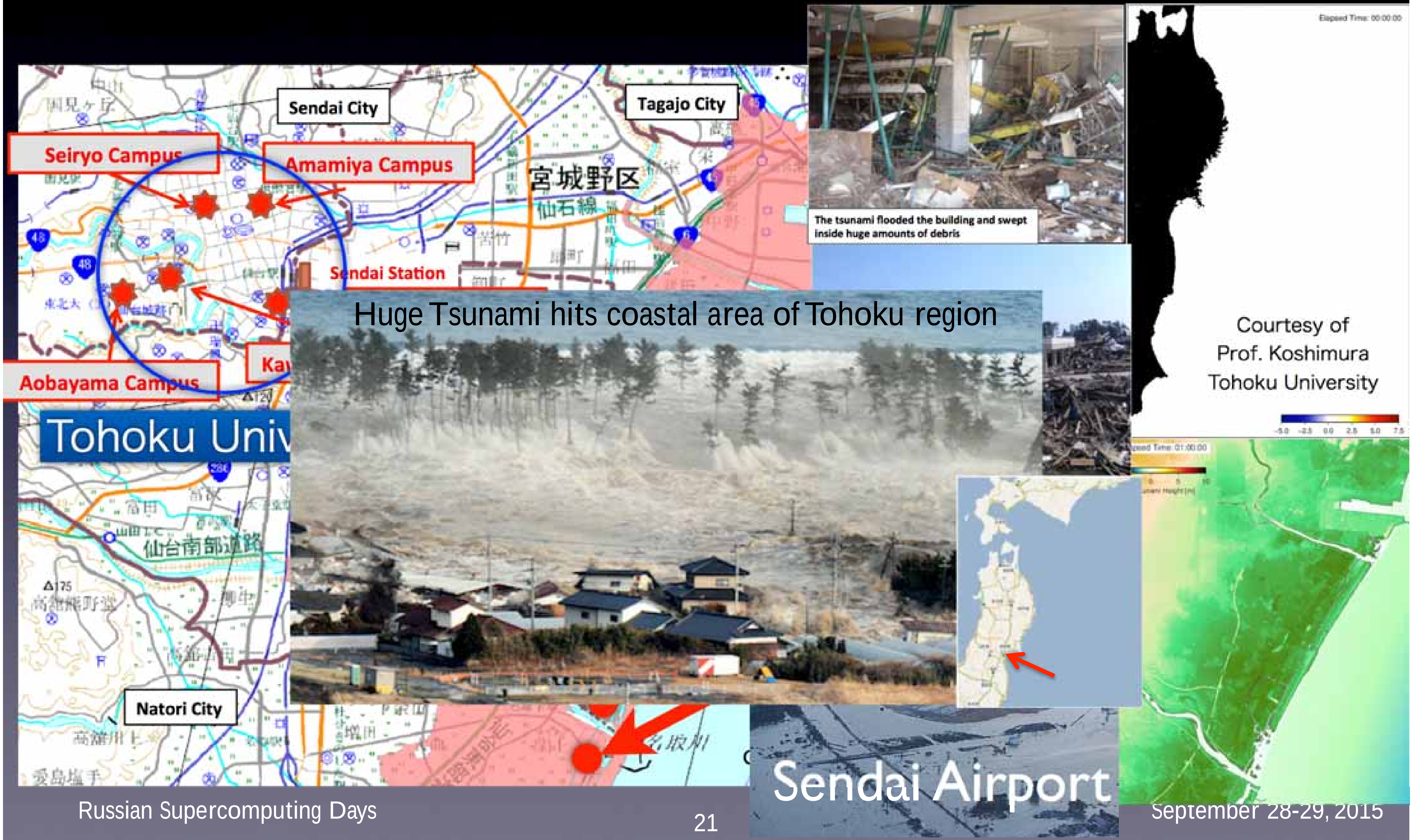


Efficiency Evaluation of HPCG Performance



A Real-Time Tsunami Inundation Forecasting System on SX-ACE for Tsunami Disaster Prevention and Mitigation

Motivation:
Serious Damage to Sendai Area Due to 2011 Tsunami Inundation



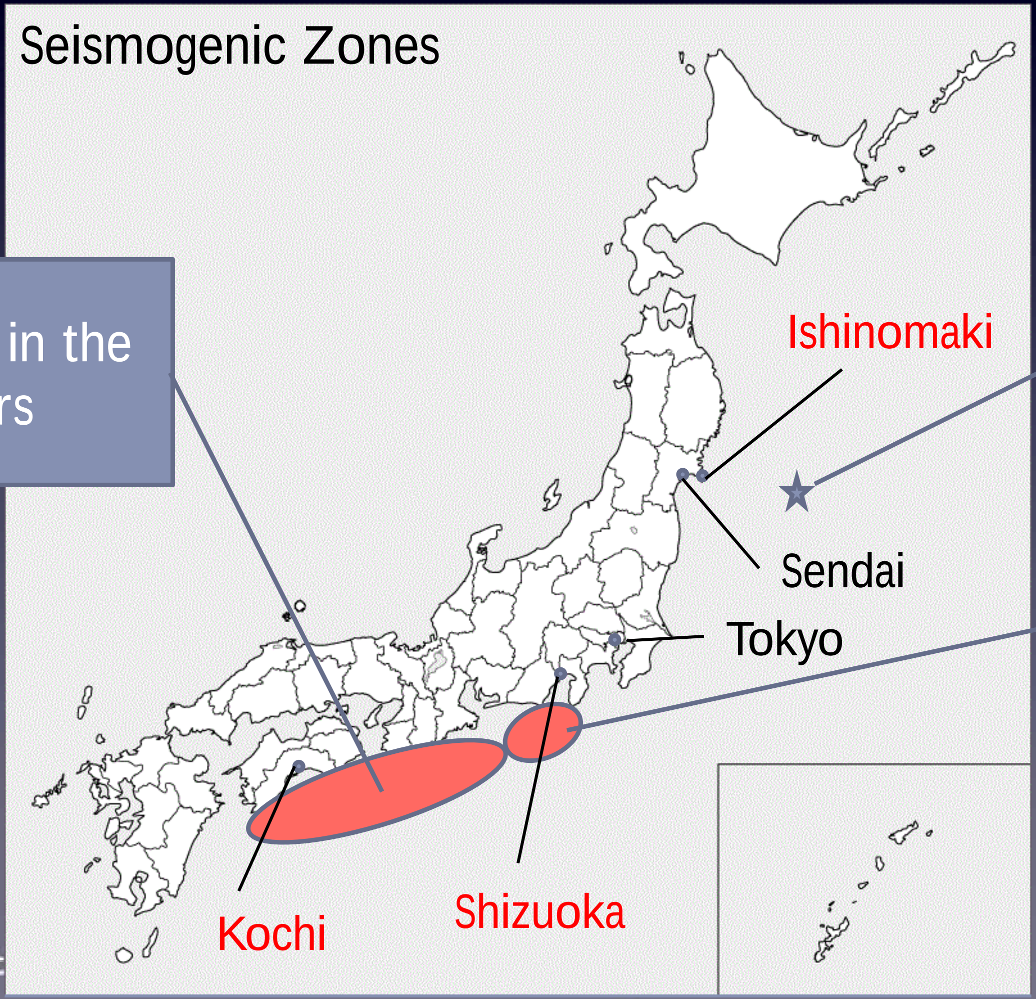
The tsunami flooded the building and swept inside huge amounts of debris

Huge Tsunami hits coastal area of Tohoku region

Courtesy of
Prof. Koshimura
Tohoku University

It's not End:
High Probability of Big Earthquakes in Japan

- Japan may be hit by severe earthquakes and large tsunamis in the next 30 years



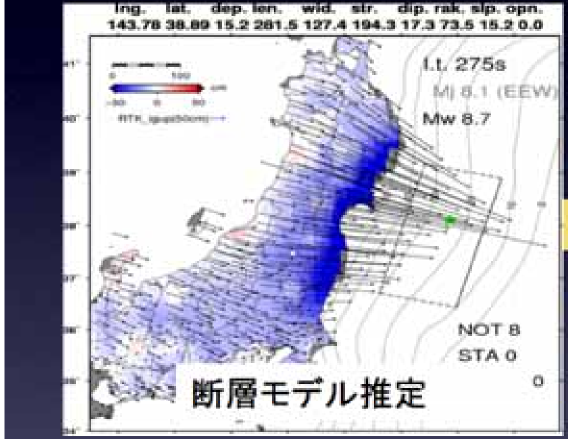
70 % probability in the next 30 years

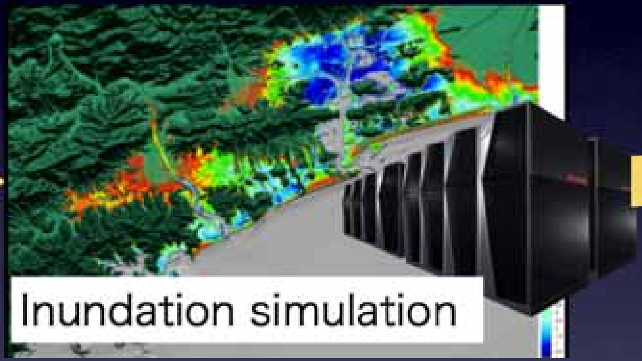
The 2011 Great Tohoku Earthquake

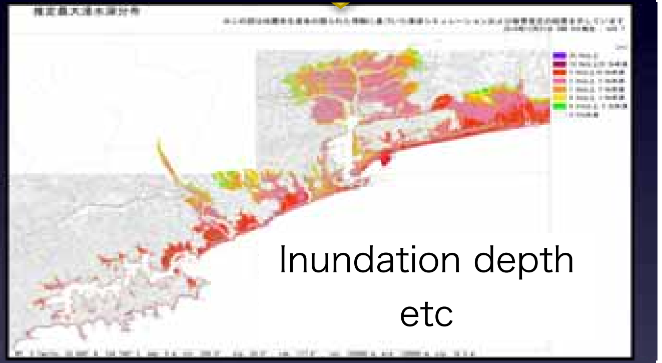
88 % probability in the next 30 years

Design and Development of A Real-Time Tsunami Inundation Forecasting System

GPS-Observation | Simulation on SX-ACE | Information Delivery







Fault estimation based on GPS data

10-m mesh models of coastal cities

Just-In-Time access of Visualized information by local governments

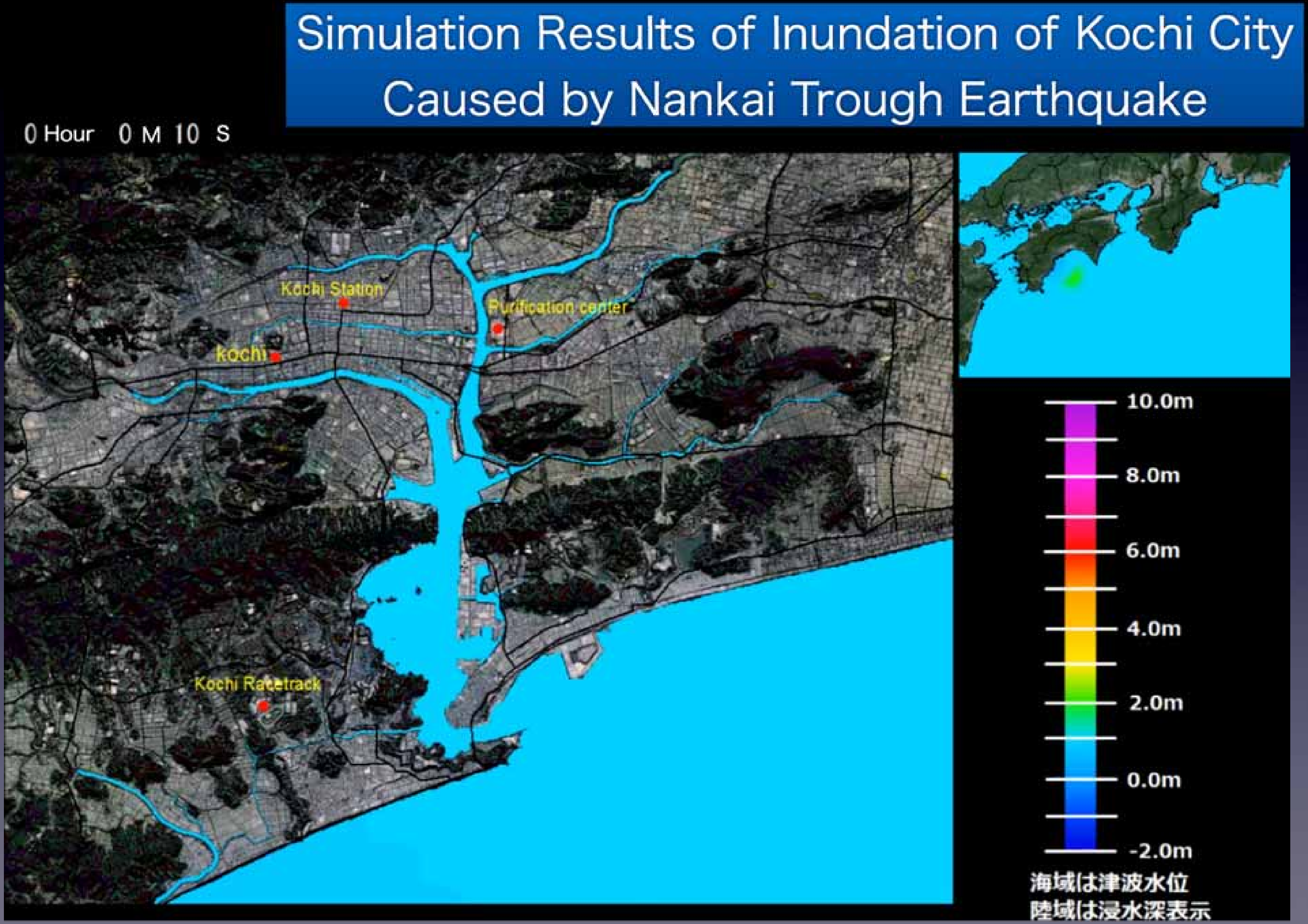
< 8 min

< 8 min

< 4 min

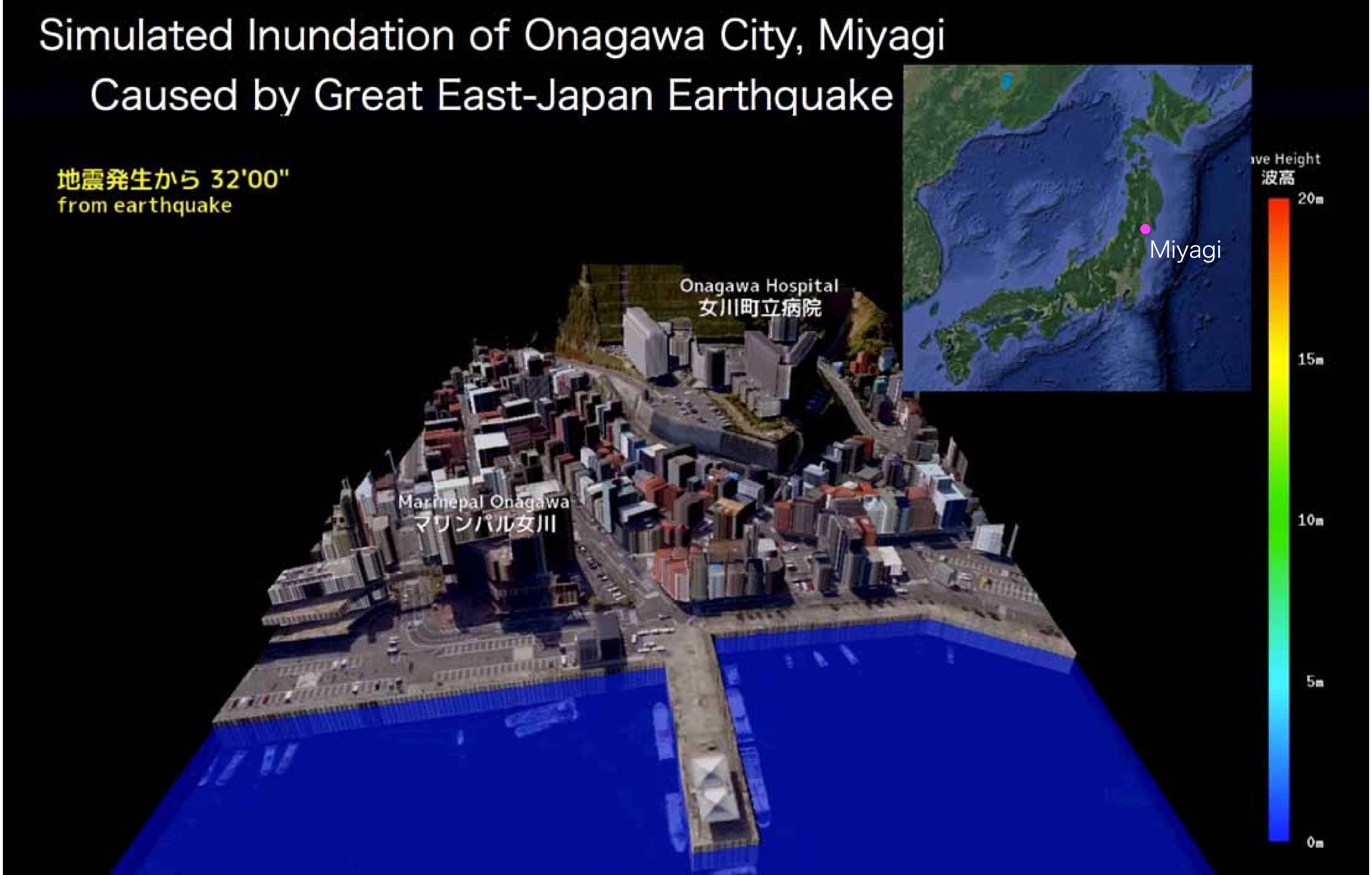
< 20 min

Demo: Visualization of Simulation Results



Simulation Results of Inundation of Kochi City Caused by Nankai Trough Earthquake

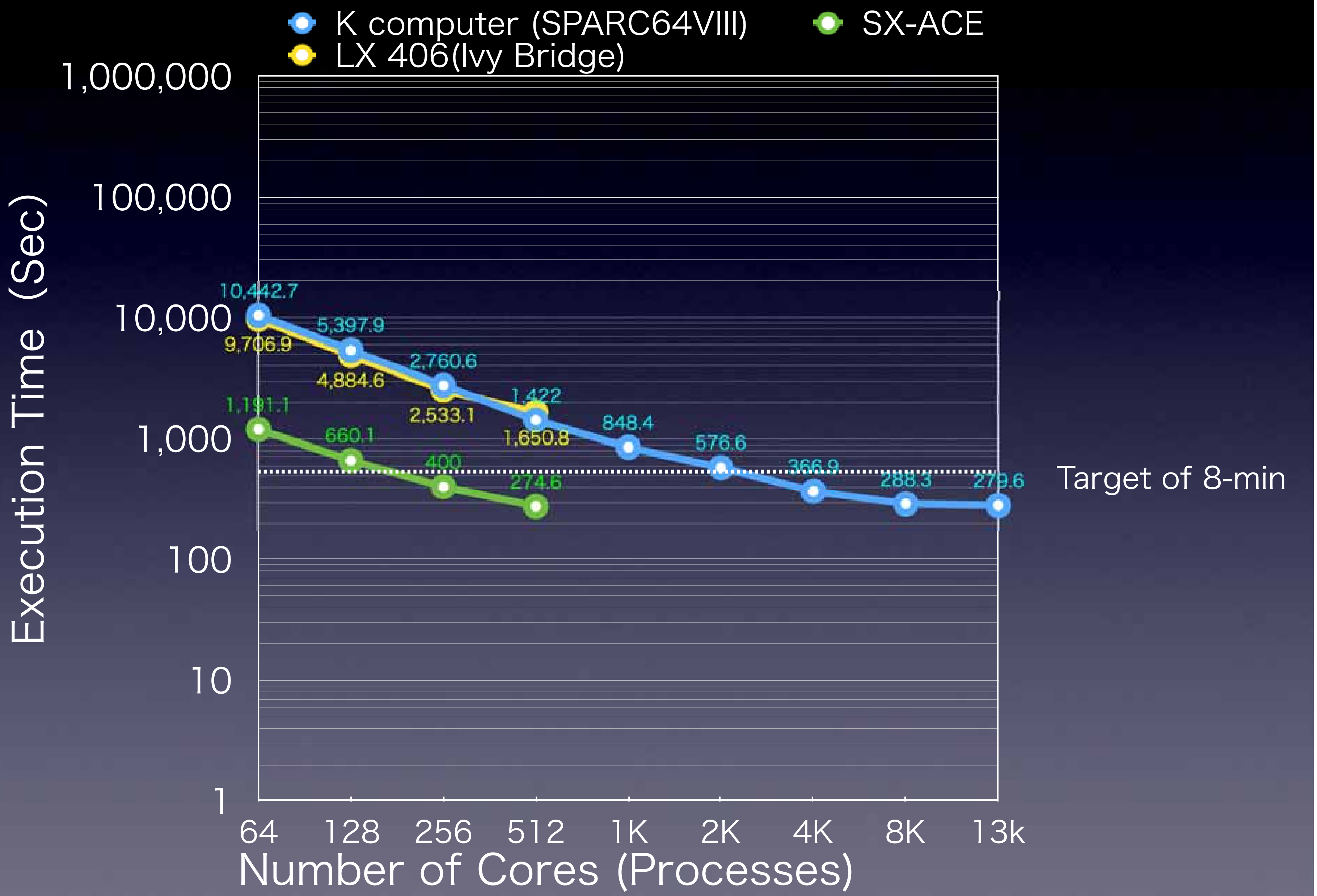
Demo:Visualization of Simulation Results



Simulated Inundation of Onagawa City, Miyagi
Caused by Great East-Japan Earthquake

地震発生から 32'00"
from earthquake

Scalability of Tunami Code



Future Vector Systems R&D*

*This work is partially conducted with NEC, but the contents do not reflect any future products of NEC

Big Problem in HPC System Design
"Brain Infarction" of Modern HEC Systems

✓Imbalance between peak flop/s rate and memory bandwidth of HEC systems result in an inefficiency in computing

- only a small portion of peak flop/s contributes to execution of practical applications in many important science and engineering fields.
- A large amount of sleeping flop/s power is wasted away!!!

➡So fa, it would be OK because Moore's low works, but now it does not make sense, as the end of Moore's law is approaching!

★So, we have to become much more smart for design of Future HEC systems,

- Because it is very hard to obtain cost reduction by Moore's law, in particular, a sky-rocketing cost increase in semiconductor integration fabrication in the eras of 20nm or finer technologies

  ✓The silicon budget for computing is also not free any more!

- Exploit sleeping flop/s efficiently by redesign/reinvent of memory subsystems to protect HEC systems from "Serious Brain Infarction"

- Use precious silicon budget (+ advanced device technologies) to effectively design mechanisms that can supply data to computing units smoothly.

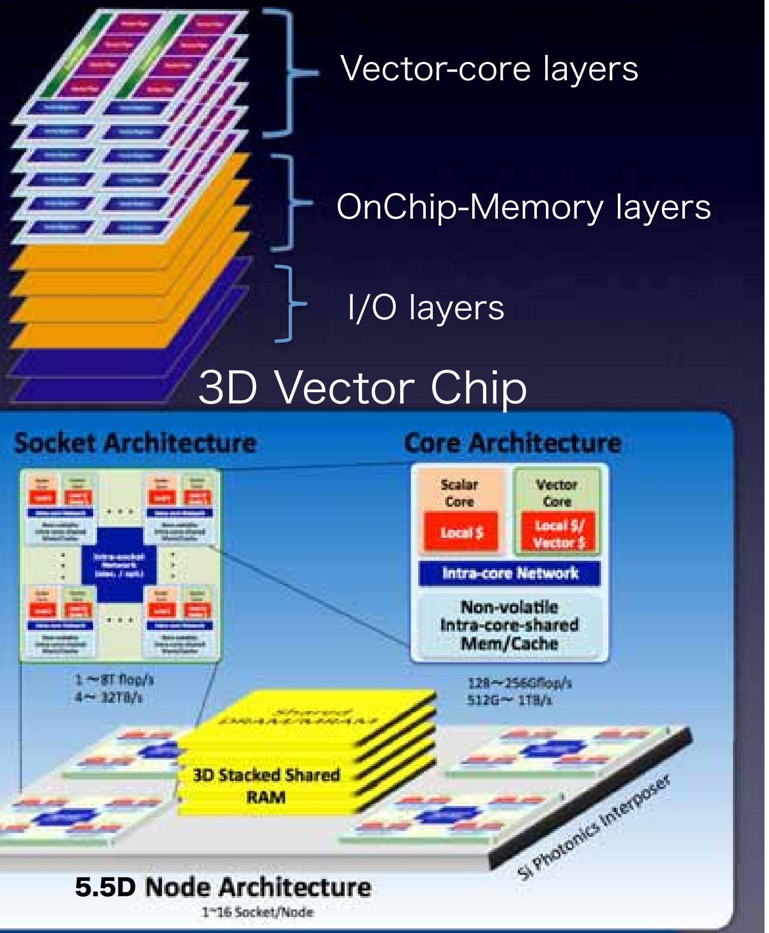
Toward the Next-Generation Vector System Design

- ★Much more focusing on sustained performance by increasing processing efficiency, not heavily depending on peak performance for the design of the next generation vector systems

- ★Architecture design of high-performance vector cores connected to an advanced memory subsystem at a high sustained bandwidth
  - find a good balance between computing performance and memory performance of future HEC systems to maximize computing efficiency of wide variety of science and engineering applications.
  - achieve a certain level of sustained performance with a moderate number of high-performance cores/nodes with a lower power budget.
    - shorter diameter, lower latency, high-BW networks also become adoptable.

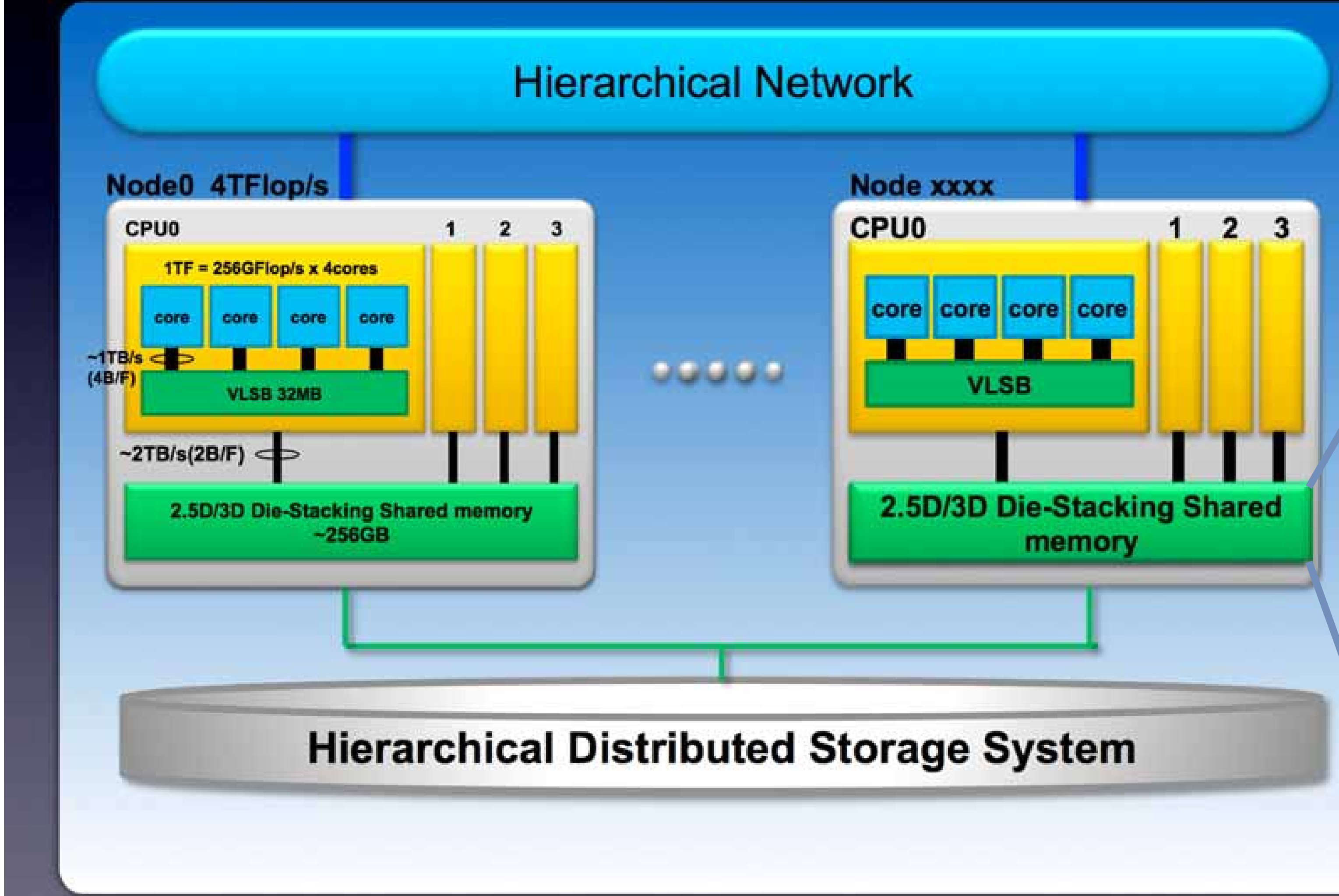
Next Generation Vector Architecture

(Kobayashi. et al 2012)

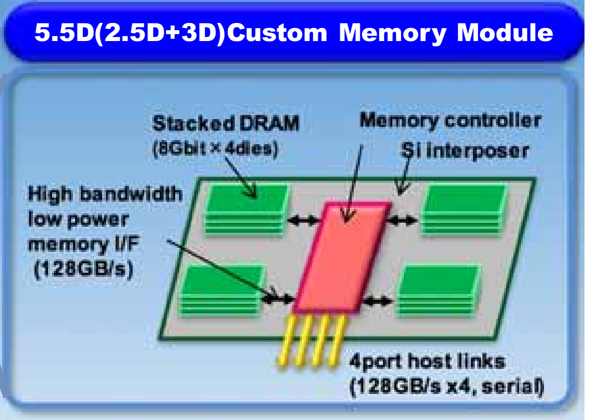


- Key technologies:
  - High throughput vector-multicore architecture
  - On-chip high BW cache & off-chip large memory
  - 5.5D (2.5D & 3D) device technologies for high-throughput computing, high memory bandwidth, and low-power consumption in the more-than-moor era

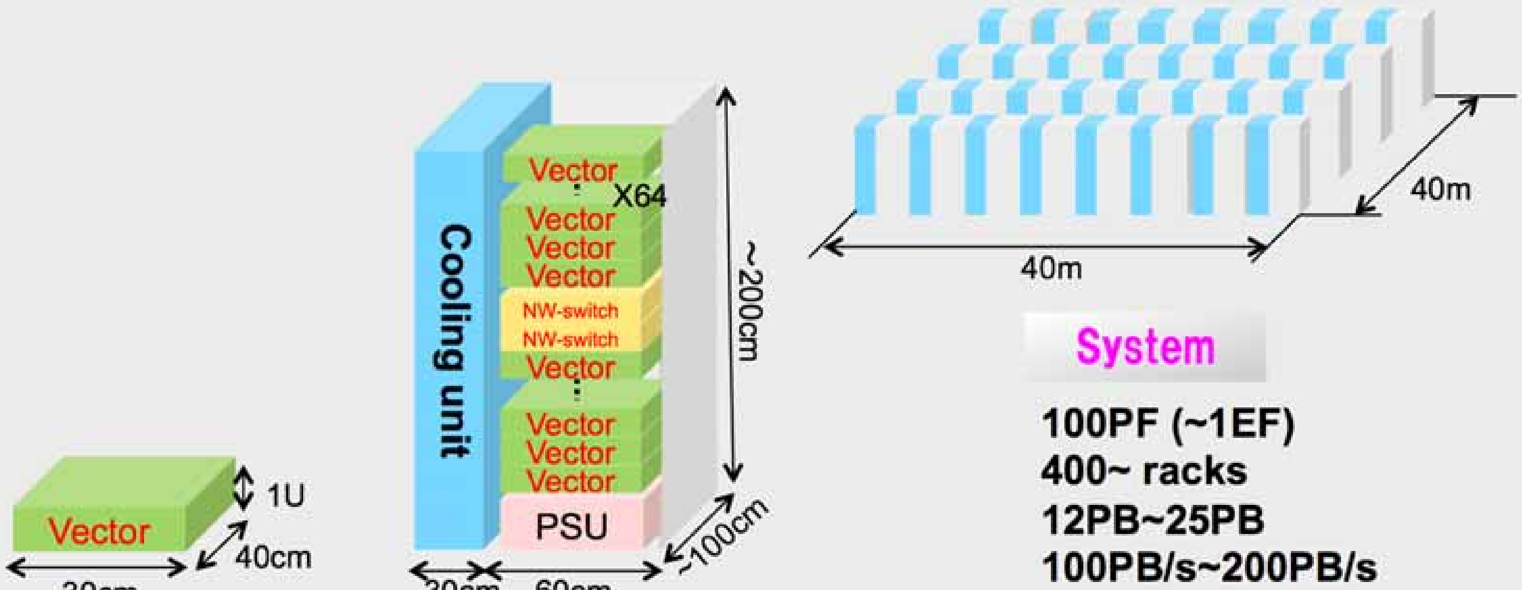
Feasibility Study of the Next-Generation Vector System Targeting Year around 2018+?


Hierarchical Network
Node0 4TFlop/s
CPU0
1
2
3
1TF = 256GFlop/s x 4cores
core
core
core
core
~1TB/s
(4B/F)
VLSB 32MB
~2TB/s(2B/F)
2.5D/3D Die-Stacking Shared memory ~256GB
Node xxxx
CPU0
1
2
3
core
core
core
core
VLSB
2.5D/3D Die-Stacking Shared memory
Hierarchical Distributed Storage System

	System Spec
Tread/proc	4
PeakPF/proc	1TF
PeakPF/100,000proc	100PF
Nodes /100,000proc	25,000
Total MemBW	200PB/s
Total Cache Capacity	3.2TB



Feasibility Study of the Next-Generation Vector System Targeting Year around 2018+?



Node

4TF
512GB~1TB
4TB/s~8TB/s
800~1000W

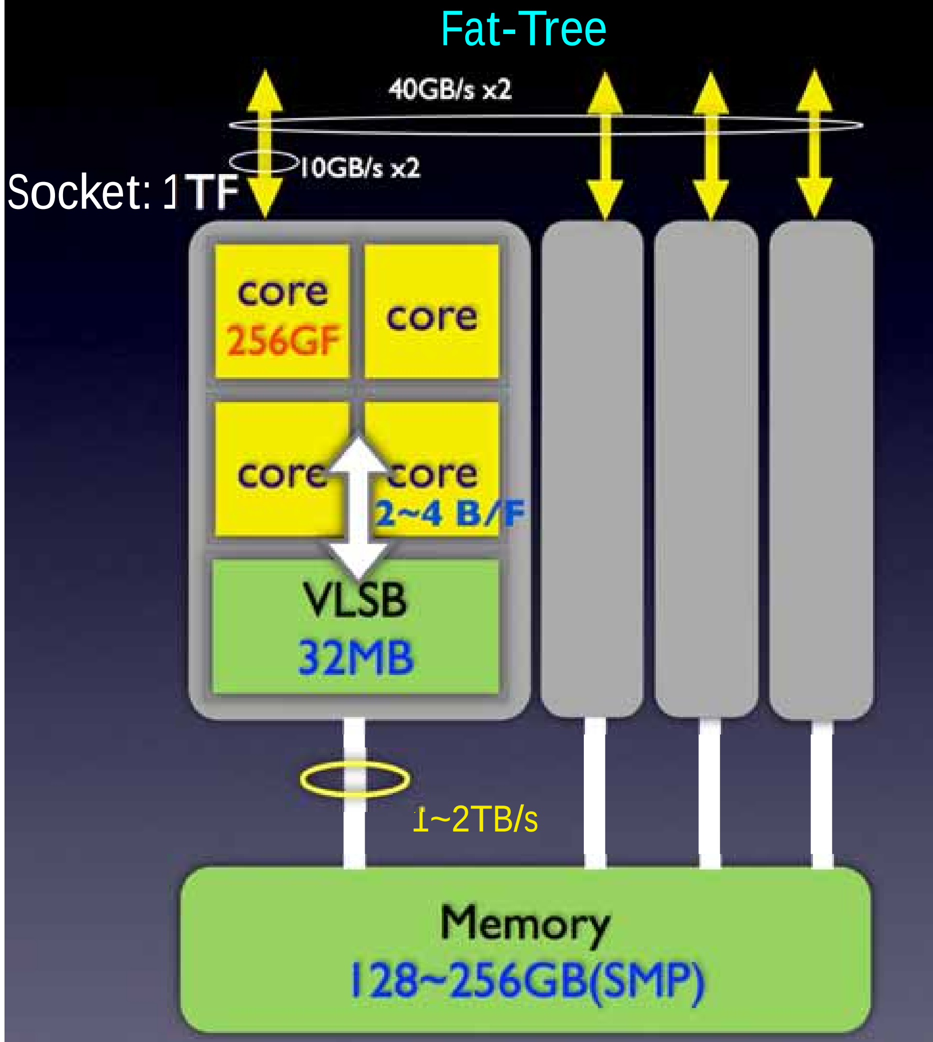
Rack

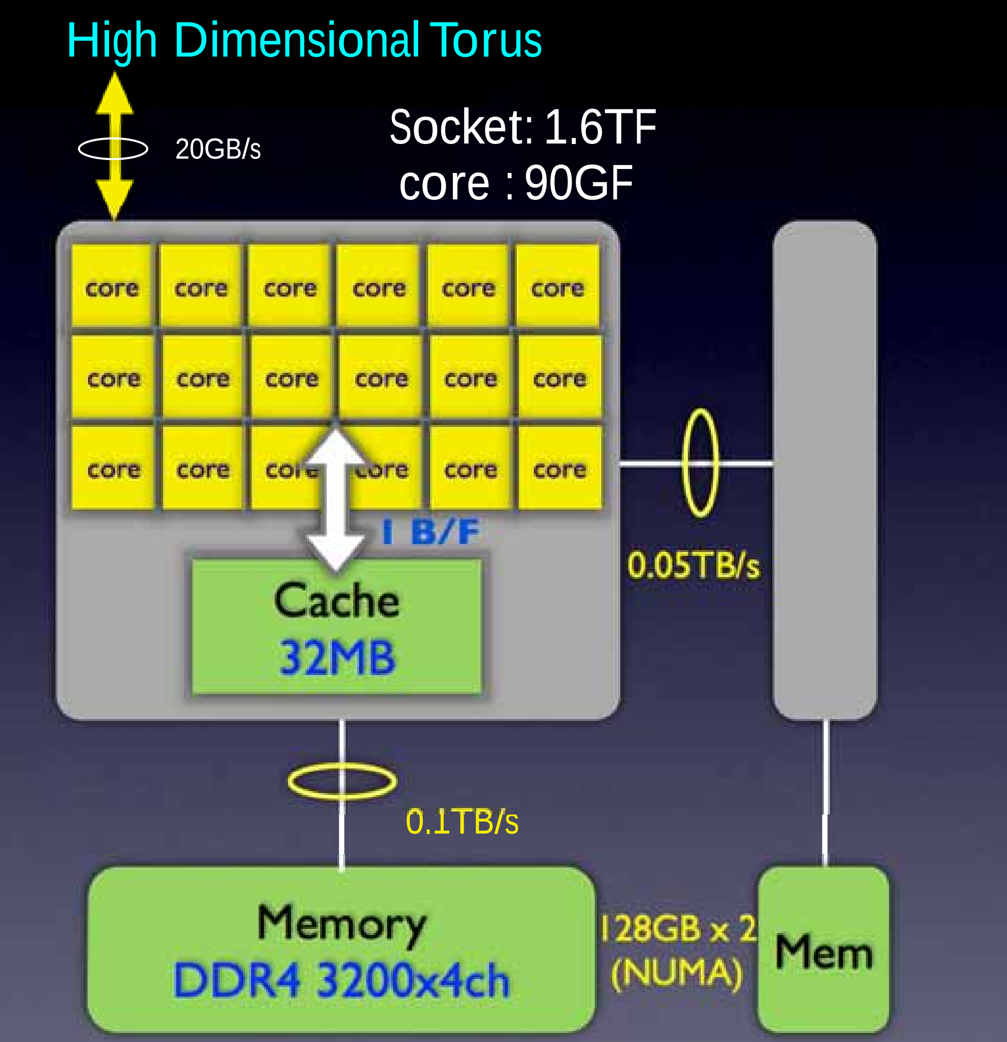
256TF(64nodes)
32TB~64TB
256TB/s~512TB/s
52~64KW

System

100PF (~1EF)
400~ racks
12PB~25PB
100PB/s~200PB/s
20~30MW

Performance Comparison of Our Target System with a Commodity-Based Future System





Our System

SMP(UMA) Architecture

4TF, ~8TB/s(~2B/F), ~512GB, 1~1.6KW

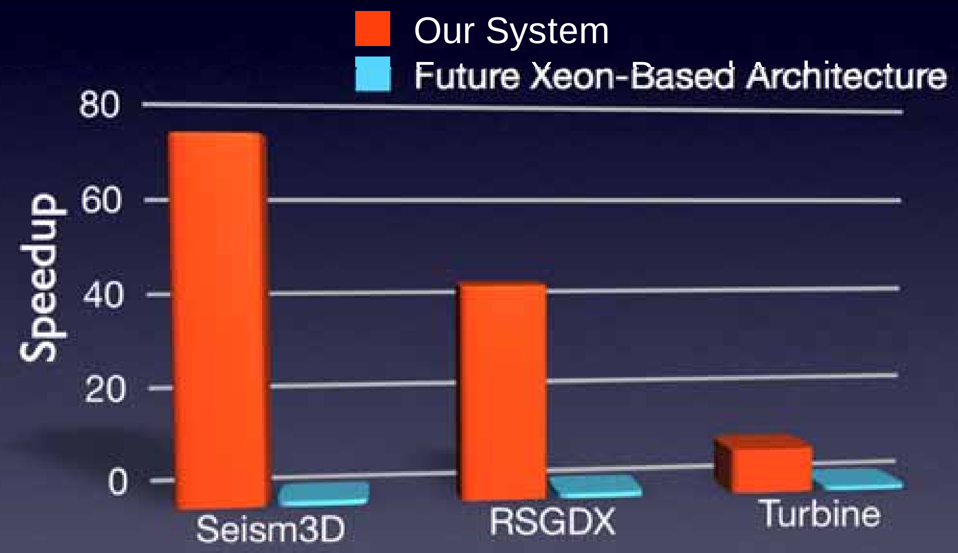
Commodity-based System

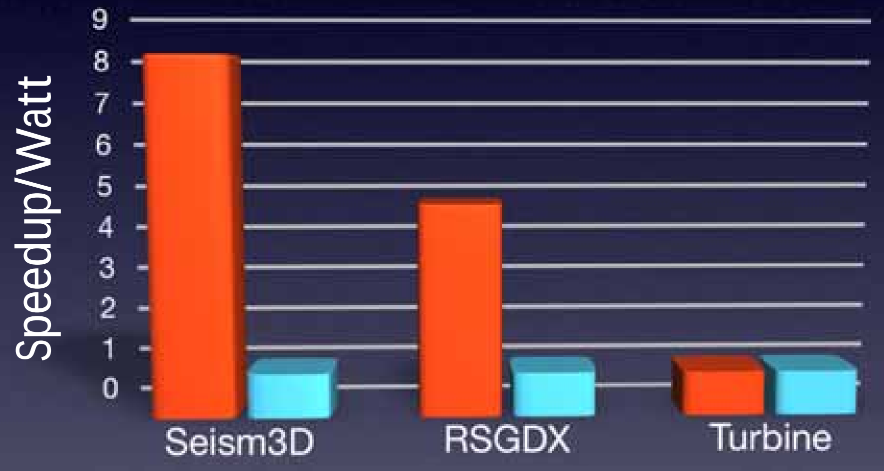
NUMA Architecture

3.3TF, 0.2TB/s(0.1B/F), 128GBx2, 0.4KW

Performance Estimation

- In the case of the same number of processes(100,000proc)
  - Performance normalized by Xeon-based System





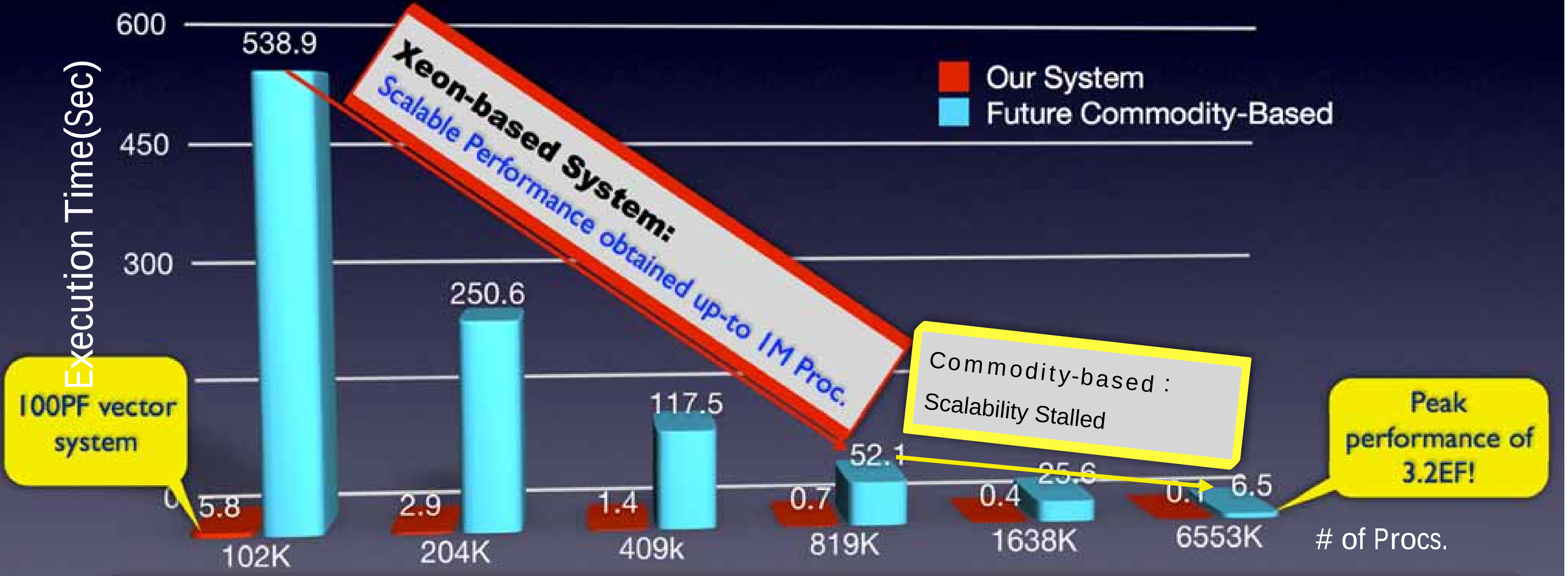
	Our System	Xeon-based	Ratio
Tread/proc	4	4	
PeakPF/proc	1TF	0.36TF	2.79
PeakPF/100,000proc	100PF	35.84PF	2.79
Nodes /100,000proc	25,000	11,111	2.25
Total MemBW	200PB/s	3.4PB/s	58.6
Total Cache Capacity	3.2TB	0.8TB	4

- 4x~8x in Performance per watt in Seism3D and RSGDX because of their high B/F requirement
- Comparable performance per watt in Turbine

Performance Estimation (2/2)

- Scalability Analysis when increasing the number of processes

Seism3D



Xeon-based System:
Scalable Performance obtained up-to 1M Proc.

Commodity-based :
Scalability Stalled

100PF vector system

Peak performance of 3.2EF!

- Xeon-based System needs 6.4M processes , which needs a peak of 3.2EF to achieve the equivalent sustained performance of our 100PF system

Summary

- ★ SX-ACE shows high sustained performance compared with SX-9, in particular short-vector processing and indirect memory accesses
- ★ Well balanced HEC systems regarding memory performance is the key to success for realizing high productivity in science and engineering simulations in the post peta-scale era
- ★ We explore the great potential of the new generation vector architecture for future HPC systems, with new device technologies such as 2.5D/3D die-stacking
  - ✳ High sustained memory BW to fuel vector function units with lower power/energy.
    - ✓ The on-chip vector load/store unit can boost the sustained memory bandwidth energy-efficiently
- ★ When such new technologies will be available as production services?
  - ✳ Design tools, fab. and markets steer the future of the technologies!

Final Remarks

Now It's Time to *Think Different!*

~Make HPC Systems Much More Comfortable and Friendly ~

- ★ Targeting HPC systems design for the entry and middle-class of HPC community in daily use, not for top, flop/s-oriented, in special use!
- Spending much more efforts/resources to exploit the potential of the system even with the moderate number of nodes and/or cores for daily use that requires high-productivity in simulation.

- Even though this approach sacrifices exa-flop/s level peak performance!
  - Seeking Exa-flops with accelerators and its downsizing deployment to entry and middle classes are NOT a smart solution in the post-peta scale era.

Let's make The Supercomputer for the Rest of US happen!